The Normal Kernel Coupler: An adaptive Markov Chain Monte Carlo method for efficiently sampling from multi-modal distributions

Gregory R. Warnes[1]

Techical Report no. 39,

Department of Statistics

University of Washington

March 1, 2001

[1]Gregory R. Warnes is a Coordinator, Biometrics and Reporting, Pfizer Global Research and Development, MS 8260-114, Eastern Point Road, Groton, CT 06340, (E-mail: gregory_r_warnes@groton.pfizer.com). This research was supported in part by NIH/NIAID Grant no. 5 T32 AI07450-09, NIH Grant no. 1 PO1 CA76466, NIH Grant no. 1 PO1 CA76466, and ONR Grant no. N00014-96-1-0192 while the author was a graduate student at the Department of Biostatistics, University of Washington. The author thanks Adrian E. Raftery, Anthony J. Rossini, and Thomas S. Lumley for helpful comments.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 MAR 2001		**00-03-2001 to 00-03-2001**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Normal Kernel Coupler: An adaptive Markov Cahin Monte Carlo method for efficiently sampling from multi-modal distributions	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**37**	

Abstract

The Normal Kernel Coupler (NKC) is an adaptive Markov Chain Monte Carlo (MCMC) method which maintains a set of current state vectors. At each iteration one state vector is updated using a density estimate formed by applying a normal kernel to the full set of states. This sampler is ergodic (irreducible, Harris recurrent and aperiodic) for any continuous distribution on d-dimensional Euclidean space. The NKC outperforms standard MCMC methods on a variety of unimodal and bimodal problems in low to moderate dimension. We illustrate the utility of the NKC by fitting a mixture model for genetic instability in cancer cells. This model, which which has two distinct and dissimilar modes, is not well handled by standard MCMC methods. In contrast, the NKC efficiently samples from this model and yields results that are consistent with current scientific understanding.

Keywords: Bayesian Estimation, Multi-chain samplers, Loss of heterozygosity, Cancer genetics, Posterior distribution

1 Introduction

Markov Chain Monte Carlo (MCMC) is a method of performing numerical integration for analytically intractable functions that can expressed as distributions (Metropolis et al., 1953; Hastings, 1970). After an initial burn-in period, a properly constructed MCMC sampler will generate (non-independent) samples from arbitrarily complicated probability distributions without requiring specification of normalizing constants. Ergodic averages of the samples thus generated can be used to estimate the expectation of arbitrary functions under the target distribution.

Although many MCMC techniques are available which effectively sample from unimodal distributions, efficient sampling from multi-modal distributions remains a difficult problem.

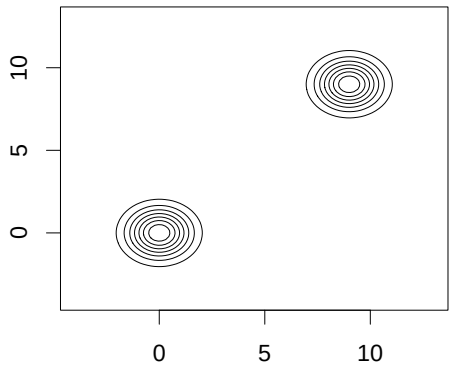
To illustrate the difficulty in using MCMC to sample from multi-modal distributions, consider a distribution formed by combining two equally weighted bivariate normals, one centered at $(0, 0)$ and the other at $(9, 9)$. This distribution can be expressed as

$$X \sim \frac{1}{2}\mathrm{N}_2\big(\,(0,0),\ \mathbf{I}_2\,\big) + \frac{1}{2}\mathrm{N}_2\big(\,(9,9),\ \mathbf{I}_2\,\big). \tag{1}$$

Contours of the corresponding density are plotted in Figure 1.

The two most frequently applied MCMC techniques are the Gibbs sampler (Geman &

Figure 1: Density contours for a mixture of 2 unit-variance bivariate normals

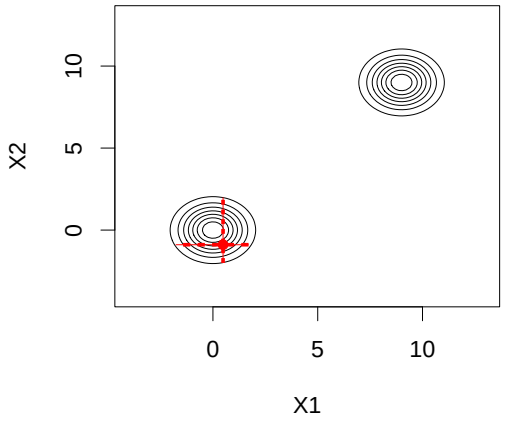


Geman, 1984) and the variable-at-a-time Metropolis sampler (Metropolis et al., 1953) with a normal proposal. Both methods update only a single parameter at each iteration, so that all moves are parallel to one of the coordinate axes. Since the modes of the example distribution are non-overlapping and do not lie along the coordinate axes, neither sampler is able to move effectively between the modes of this distribution (see Figure 3(a) and 3(b)). While it is possible in this case to transform the parameters so that the modes lie along the axes, finding a suitable transformation may be difficult or impossible for realistic problems.

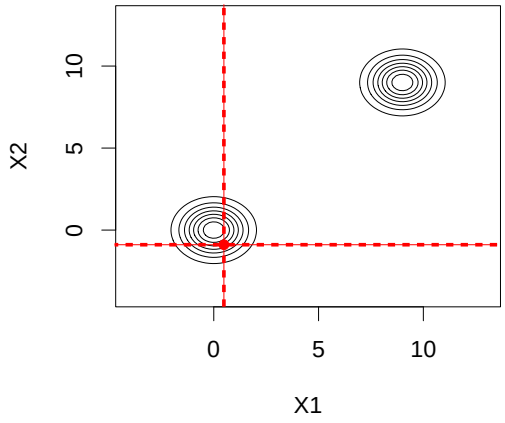
Another standard MCMC technique is the random walk Metropolis sampler using a multivariate normal proposal distribution. Unlike the variable-at-a-time methods, this Metropolis sampler can generate moves between the modes of the example distribution, provided that the variance of the proposal distribution is sufficiently large. Unfortunately, a proposal with a variance large enough to ensure moves between the modes will also generate many points that are near neither mode (see Figure 3(c)). This will cause a large proportion of candidate points to be rejected, and will make the sampler very inefficient. This is particularly problematic in high dimensions and when the modes are well separated.

Although the MCMC literature contains numerous references to the difficulties created by multi-modal distributions, there are only a handful of MCMC techniques designed to effectively sample from such distributions. Gelman & Rubin (1992) recommended creating a custom independence proposal constructed from a mixture of multivariate normal distributions based on pre-simulation exploration. Geyer (1991) introduced Metropolis-Coupled MCMC, which uses a set of concurrent MCMC samplers each operating on one of a set of successively smoother distributions. Coupling these samplers by occasionally swapping current states allows the roughest distribution, corresponding to the density of interest, to inherit mobility possible in the smoother distributions. Meanwhile Mariani & Parisi (1992) and later Geyer & Thompson (1994) describe Simulated Tempering, a related method which allows a single sampler to move through the set of distributions rather than having a set of concurrent samplers. Neal (1996) extended this work to reduce the effort required for tuning. Recently, Tjielmeland & Hegstad (2000) introduced a method for incorporating "mode

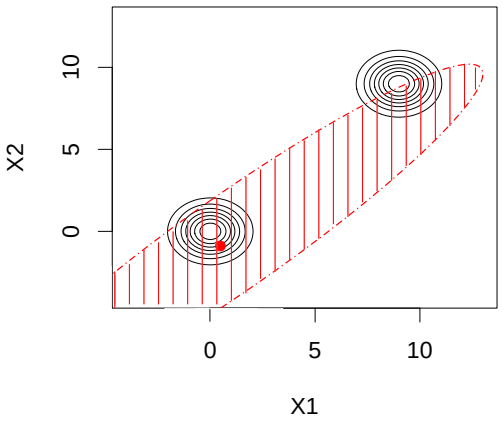
Figure 2: Possible moves for the Gibbs Sampler, variable-at-a-time Metropolis, and (multivariate) random-walk Metropolis for the point represented by the red dot.



(a) Gibbs Sampler



(b) Variable-at-a-time Metropolis



(c) Random Walk Metropolis

finding" moves using a standard numerical maximization technique with a random starting locations into a standard MCMC sampler.

Unfortunately, each of these methods requires considerable problem-specific effort to be invested before useful results can be obtained. The construction of a custom proposal distribution requires not only the location of the modes of the distribution, but also requires determination of the appropriate covariance matrix for each mode. The tempering methods require selection of the smoothed distributions and derivation of appropriate probabilities of swapping between samplers. In addition, only a small fraction of the MCMC states generated by the tempering methods correspond to the distribution of interest, reducing their efficiency.

There remains a need for MCMC methods that efficiently sample from multi-modal distributions and that minimize the amount of effort required from the statistician. It is in this context that we introduce the Normal Kernel Coupler (NKC).

In section 2, we introduce the Normal Kernel Coupler. Section 3 discuss the convergence properties of the NKC. We report on a simulation study comparing the performance of the NKC with several standard methods in section 4. Section 5 shows the successful application of the NKC to a bimodal model for the genetic instability of esophageal cancers. We provide conclusions and discussion in section 6.

2 The Normal Kernel Coupler (NKC)

The Normal Kernel Coupler (NKC) is an MCMC sampler that maintains a set of current states, each of which converges to the same target distribution. At each iteration, a new value is proposed for one component state using a kernel density estimate constructed from the entire current set. Since the kernel density estimate makes very few assumptions about the form of the target distribution, the Normal Kernel Coupler's efficiency is largely independent of the number and location of modes. When properly constructed, the NKC efficiently samples from both unimodal and multi-modal target distributions, even when the distribution is oddly shaped or the modes are well separated.

2.1 The Algorithm

Let $\pi(X)$, defined on $X \in \Re^d$ (d-dimensional Euclidean space), be the density of the distribution of interest. Let $p(X)$ be a function which is proportional to $\pi(X)$. Let $X_t^{(\cdot)} = (X_t^{(1)}, \ldots, X_t^{(C)})$, be a vector of component states where each $X_t^{(i)} \in \Re^d$. Note that subscripts index time, while parenthesized superscripts index component states.

We will use $\mathrm{N}_d\big(\,\mu,\ \Xi\,\big)$ to represent a d-variate normal distribution with mean vector μ and covariance matrix Ξ. In a slight abuse of notation we will use $\mathrm{N}_d\big(\,v\,\big|\,\mu,\ \Xi\,\big)$ to represent the *density* at v of a d-variate normal with mean vector μ and covariance matrix Ξ.

The Normal Kernel Coupler iterates through six step a variant of the Metropolis-Hastings update cycle:

1. **Select** a component state, $X_t^{(i)}$ $i \in 1, .., C$, to update.

2. **Propose** a new state $Y^{(i)}$ for component i using a normal kernel density estimate by randomly selecting a source component, $X_t^{(u)}$, where

$$u \sim \text{Discrete Unifiorm}(1, .., C)$$

and then generating a value from a normal centered at $X_t^{(u)}$:

$$Y^{(i)}|u \sim \mathrm{N}_d\big(\,X_t^{(u)},\ h^2\mathbf{V}\,\big),$$

so that the density of $Y^{(i)}|X_t^{(\cdot)}$ is

$$\begin{aligned} q_k(Y^{(i)}|X_t^{(\cdot)}) &= q_k(Y^{(i)}|X^{(i)}, X^{(-i)}) \\ &= \frac{1}{C}\sum_{j\neq i}^{C} \mathrm{N}_d\big(\,Y^{(i)}|X_t^{(j)},\ h^2\mathbf{V}\,\big) + \frac{1}{C}\mathrm{N}_d\big(\,Y^{(i)}|X_t^{(i)},\ h^2\mathbf{V}\,\big) \end{aligned}$$

where h^2 is a bandwidth tuning constant and $\mathbf{V}$ determines the shape and scale of the normal kernel.

3. **Compute** the Metropolis-Hastings acceptance probability

$$\begin{aligned}\alpha(X_t^{(i)}, Y^{(i)}|X_t^{(-i)}) &= \min\left\{1, \frac{\pi(Y^{(i)})\ q_k(X_t^{(i)}|Y^{(i)}, X_t^{(-i)})}{\pi(X^{(i)})\ q_k(Y^{(i)}|X_t^{(i)}, X_t^{(-i)})}\right\} \\ &= \min\left\{1, \frac{p(Y^{(i)})\ q_k(X_t^{(i)}|Y^{(i)}, X_t^{(-i)})}{p(X^{(i)})\ q_k(Y^{(i)}|X_t^{(i)}, X_t^{(-i)})}\right\}.\end{aligned}$$

4. **Accept** the proposed point $Y^{(i)}$ and set

$$X_{t+1}^{(i)} \leftarrow Y^{(i)}$$

with probability $\alpha(X_t^{(i)}, Y^{(i)}|X_t^{(-i)})$, otherwise,

Reject the proposed point and set

$$X_{t+1}^{(i)} \leftarrow X_t^{(i)}.$$

5. **Copy** the remaining states

$$X_{t+1}^{(-i)} \leftarrow X_t^{(-i)}$$

6. **Increment** time: $t \leftarrow (t+1)$.

3 Convergence

A concern with adaptive MCMC methods is the possibility that such methods may fail to converge to the desired stationary distribution. Fortunately, the NKC lends itself to a straightforward proof that it is ergodic (irreducible, Harris recurrent and aperiodic), which is sufficient to ensure convergence.

Define the joint state as the vector formed by concatenating each of the component states,

$X^{(\cdot)} = (X^{(1)}, .., X^{(C)})$, the joint target distribution by

$$\pi_*(X^{(\cdot)}) = \prod_{i=1}^{C} \pi(X^{(i)})$$

and the joint proposal density by

$$q_*(Y^{(\cdot)}|X^{(\cdot)}) = q_k(Y^{(c_t)}|X^{(\cdot)}) \prod_{i=1}^{C} \delta(Y^{(i)} = X^{(i)})$$

where $\delta(\cdot)$ is the indicator function that takes value 1 when its argument is true and 0 otherwise, and c_t cycles through a permutation of the integers $1, \ldots, C$. With these definitions, the NKC is seen to be a variable-at-a-time Metropolis-Hastings sampler for the joint target where each component state, $X^{(i)}$, is considered a parameter. In this context, it is straightforward to prove ergodic convergence, as expressed by the following theorem (see the Appendix for the proof) :

Theorem 1 *If π is a continuous distribution on $\Re^d$, $h^2 > 0$, and $\mathbf{V}$ is positive definite, then the NKC constructed for π using h^2 and $\mathbf{V}$ is ergodic (π_*-irreducible, Harris-recurrent, and aperiodic) with unique invariant distribution π_*.*

The behavior of the individual components is then a trivial extension:

Corollary 2 *The distribution of the sequence of values taken by each component state $\mathbf{X^{(i)}} = \{X_t^{(i)}, t = 0, 1, ..\}$ converges to π.*

4 Simulation Study

We performed a simulation study to compare the efficiency of the NKC to that of three standard MCMC methods; a custom independence proposal, a variable-at-a-time Metropolis sampler, and a random-walk Metropolis sampler using a multivariate normal proposal.

4.1 Target Distributions

For the simulations, we constructed a set of seven test distributions which abstract different characteristics of posterior densities encountered in practice. Each distribution was given a descriptive title. These are "OneMode", "Narrow", "TwoMode", "BigAndSmall", "HeavyAndLight", "Banana", and "TwoNarrow". Formulae and contour plots for these distributions are given in tables 1 and 3 respectively.

OneMode This target distribution, a spherical d-variate normal with identity covariance matrix, represents an ideal target distribution for which all MCMC methods should perform well.

Narrow The second sampler is a d-variate normal with an AR-1 style covariance matrix (see table 2) with with correlation coefficient $\rho = 0.95$. It represents a more realistic target distribution which is approximately normal but which has highly correlated parameters.

TwoMode The third distribution is composed of two equally weighted d-variate normals, each with an identity covariance matrix. The first is centered at $(0, 0, .., 0)$, while the second is centered at $(9, 9, 9, ..., 9)$. This target mimics the behavior of distributions with highly separated modes, where each mode is essentially identical and is well modeled by a multivariate normal.

BigAndSmall The fourth sampler is also formed using two equally weighted d-variate normals, one at $(0, 0, .., 0)$ and the other at $(9, 9, .., 9)$. However, the second mode now has a identity covariance matrix scaled down by a factor of $\frac{1}{16}$. This target exhibits different scales and different sized basins of attraction. These features can cause samplers to incorrectly assign extra mass to the larger mode.

HeavyAndLight The fifth sampler is constructed using the same normals as the previous sampler, "BigAndSmall", but the larger mode is now assigned only $\frac{1}{8}$th of the mass.

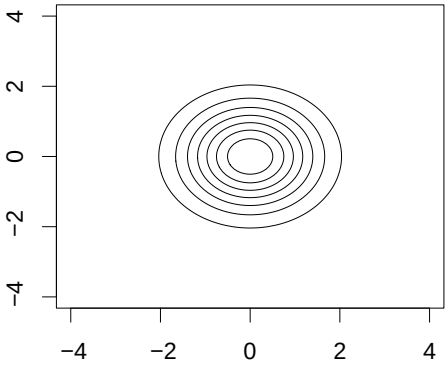
Table 1: Densities for the target distributions used in the simulation study. I is the Identity Matrix. $\mathrm{AR}_1(\rho)$ is an AR-1 covariance matrix with correlation parameter ρ (see table 2).

Target	Distribution
OneMode	$\mathrm{N}_d\big((0,0,..0),\ \mathbf{I}\big)$
Narrow	$\mathrm{N}_d\big((0,0,..0),\ \mathrm{AR}_1(0.95)\big)$
TwoMode	$\frac{1}{2}\mathrm{N}_d\big((0,0,..0),\ \mathbf{I}\big) + \frac{1}{2}\mathrm{N}_d\big((9,9,..,9),\ \mathbf{I}\big)$
BigAndSmall	$\frac{1}{2}\mathrm{N}_d\big((0,0,..0),\ \mathbf{I}\big) + \frac{1}{2}\mathrm{N}_d\big((9,9,..,9,\ \frac{1}{16}\mathbf{I}\big)$
HeavyAndLight	$\frac{1}{8}\mathrm{N}_d\big((0,0,..0),\ \mathbf{I}\big) + \frac{7}{8}\mathrm{N}_d\big((9,9,..,9),\ \frac{1}{16}\mathbf{I}\big)$
Banana	$\frac{1}{2}\mathrm{N}_d\big((-1.5,1.5,..1.5),\ \mathrm{AR}_1(-0.95)\big) + \frac{1}{2}\mathrm{N}_d\big((1.5,1.5,..,1.5),\ \mathrm{AR}_1(0.95)\big)$
TwoNarrow	$\frac{1}{2}\mathrm{N}_d\big((0,0,..,0),\ \mathrm{AR}_1(-0.95)\big) + \frac{1}{2}\mathrm{N}_d\big((9,9,..,9),\ \mathrm{AR}_1(0.95)\big)$

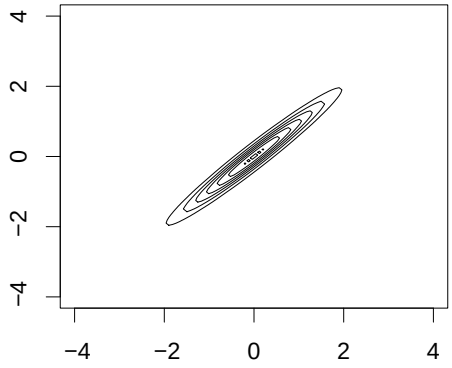
Table 2: AR-1 style covariance matrix with correlation parameter ρ

$$\mathrm{AR}_1(\rho) = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{C-1} \\ \rho & 1 & \rho & \cdots & \rho^{C-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{C-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{C-1} & \rho^{C-2} & \rho^{C-3} & \cdots & 1 \end{pmatrix}$$

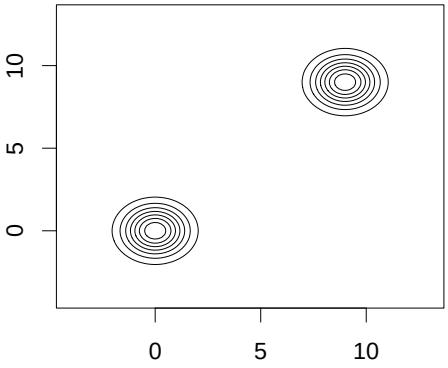
Figure 3: Contour plots of the target distributions used in the simulation study



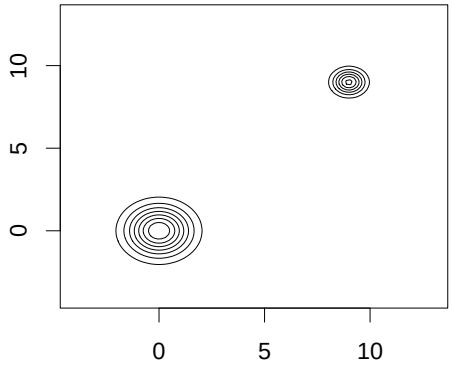
(a) OneMode



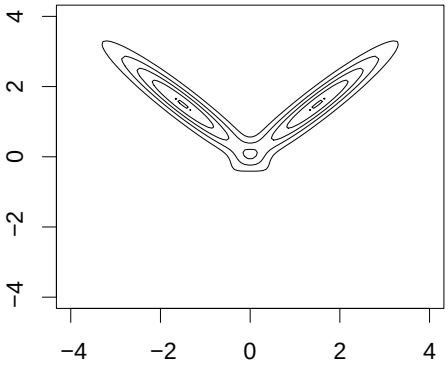
(b) Narrow



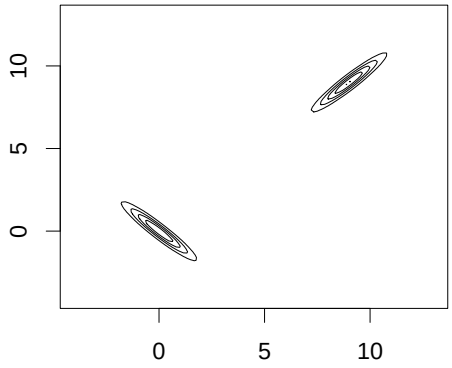
(c) TwoMode



(d) BigAndSmall (equal weight) and HeavyAndLight (weight=$\frac{1}{8}$, $\frac{7}{8}$)



(e) Banana



(f) TwoNarrow

This makes it very likely that samplers will incorrectly assign too much mass to the larger mode.

Banana The sixth target distribution is also formed from two d-variate normals. These two normals are centered at $(1.5, 1.5, .., 1.5)$ and $(-1.5, 1.5, .., 1.5)$. The first normal has an AR-1 style covariance matrix with correlation coefficient $\rho = 0.95$. The second mode has an AR-1 structure with correlation coefficient $\rho = -0.95$. This creates two narrow normals aligned perpendicularly, with some overlap at $(0, 0, .., 0)$. This target mimics the complex topologies which can can seriously reduce the efficiency of many samplers.

TwoNarrow The seventh target also has two perpendicular modes with AR-1 structure with $\rho = 0.95$ and $\rho = -0.95$ respectively. This time, the modes are quite separated, with one at $(0, 0, .., 0)$ and the other at $(9, 9, .., 9)$. This combines the problems of different covariance structure within modes with those of separated modes.

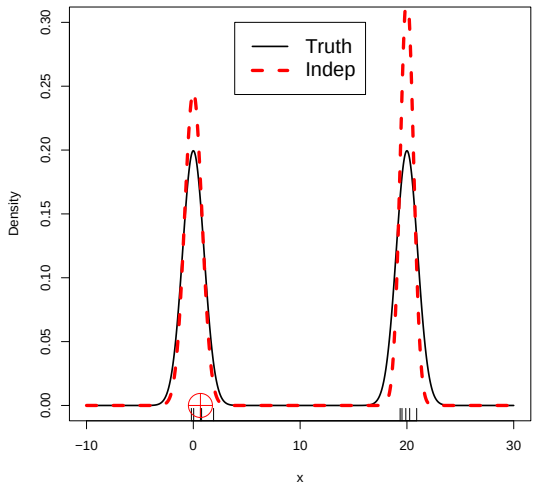
4.2 Samplers

Four samplers were used for the simulations, a Metropolis-Hastings sampler using a custom independence proposal constructed from a mixture of normal distributions, a variable-at-a time Metropolis sampler using a normal proposal, a random walk Metropolis sampler using a multivariate normal proposal distribution, and Normal Kernel Coupler.

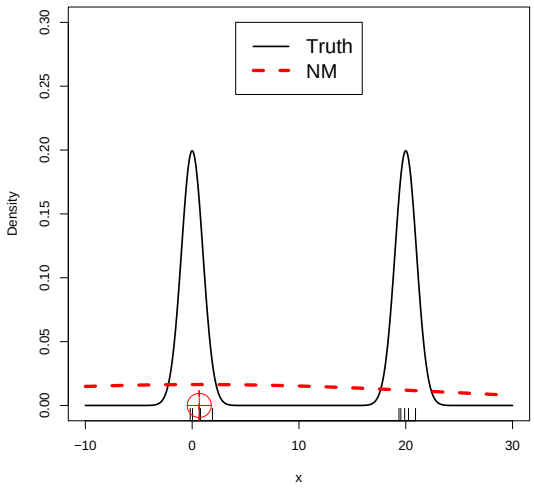
We used two versions of each sampler for the simulations. To capture the best possible performance of the individual samplers we constructed the first version using an appropriate function of the true covariance. The second sampler was constructed "adaptively" using a multi-stage tuning method method.

Our multi-stage tuning method similar to the one proposed by Raftery & Lewis (1996). A sequence of runs, each of length 1,000, was used to estimate the overall and per-mode variance. For the first run, the MCMC samplers were constructed using a preliminary variance estimate. The initial estimates of the variances of each mode was $0.5\,\mathbf{I}_d$. For unimodal distributions, this was also the the estimate of the overall variance. For bimodal

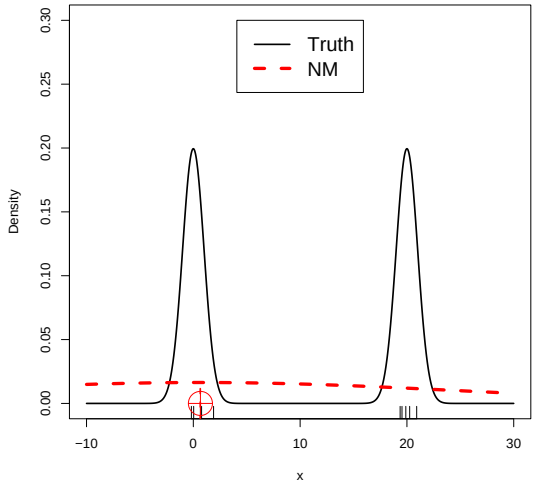
Figure 4: Proposal distributions for the samplers used in the simulation study. Ten points (tick marks along the x-axis) were sampled from a bimodal posterior (solid curve). The broken curve shows the proposal distribution for one of the points, which is marked by the circle. The variance for the proposal distributions and the density estimate for the NKC were computed using the 10 points shown.



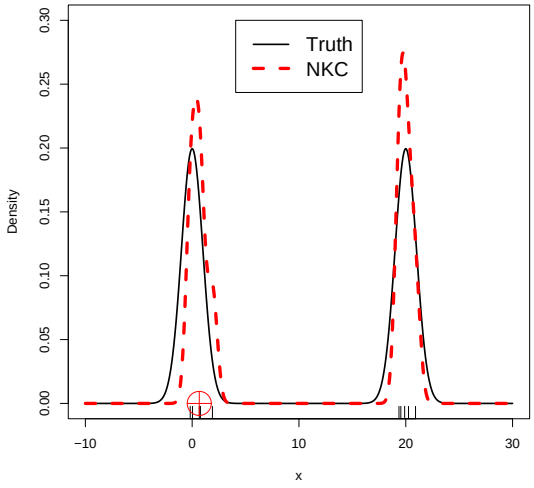
(a) Indep



(b) variable-at-a-time Metropolis



(c) random walk Metropolis



(d) NKC

distributions, the overall variance estimate was computed using the estimated the mode variance and known locations of the mode centers.

After each 1,000 iterations, new variance estimates were computed. These estimates were then used to construct the samplers for the next set of iterations. This iterative tuning process was performed for one third of the total number of iterations. EG, when the total number of iterations was 10,000, three runs of length 1,000 were used to tune the variance estimates. Once the tuning phase had been completed, the variance remained fixed for the remaining iterations.

At each stage, the overall variance was estimated by computing the variance matrix of the simulation output. For the bimodal targets, the variance of each mode was estimated by dividing the parameter space along the first dimension halfway between the modes. The sample values in each half-space were then used to estimate the corresponding mode variance.

The custom independence proposal was constructed from either a single multivariate normal distribution for unimodal targets, or from a mixture of two equally weighted multivariate normal distributions for bimodal targets. Each normal was centered at one of the posterior modes and was assigned either the true or estimated variance of the mode. Note that the independence proposal using the true variance of the modes simulates directly from the true target distribution in every case except the HeavyAndLight target. For the HeavyAndLight target the independence proposal gives equal weight to the two modes when the true distribution gives the smaller (in area) mode seven times more weight. Thus, except in this one case, this independence proposal gives the best possible performance for a sampler which maintains only a single current state.

The random-walk Metropolis sampler was constructed using a d-variate normal distribution. The variance was set to the known or estimated overall variance scaled by $\frac{2.38^2}{d}$, the optimal scaling factor derived by Gelman et al. (1995). The componentwise random-walk Metropolis sampler used a the univariate normal proposal with variance set 2.38^2 times the true or estimated marginal variance of the parameter being updated.

The NKC was constructed as described in section 2.1 with the variance matrix $\mathbf{V}$ set to

the average of the true or estimated variances of the individual modes. The scaling factor h^2 was set to $1.4\left(\frac{1}{C}\right)^{\frac{2}{4+d}}$.

4.3 Setup

Two sets of simulations were run. The first set used target distributions defined in four dimensions, and the second set used distributions defined in twenty dimensions. For both sets of simulations, the NKC used 200 component states. Twenty trials were performed for each combination of sampler, target distribution, and number of components.

The states of each sampler were randomly initialized to one of the modes of the target distribution plus a small random displacement: $X_0^{(c)} \sim N_d(\mu_i, 0.05\mathbf{I})$, where μ_i is the peak of the ith mode. This mimics the practice of initializing the MCMC samplers using modes located via a numerical maximization technique.

Since the primary cost of most MCMC simulations is the expense of evaluating the (unnormalized) density of the target distribution, the results are displayed in terms of the total number of likelihood evaluations rather than in execution time. For the four dimensional simulation, cumulative univariate means and quantiles were computed after 1,000, 2,000, 3,000, 4,000, 8,000, and 10,000 iterations. For the twenty dimensional simulation we computed cumulative univariate means and quantiles after 10,000, 20,000, 30,000, 40,000, 80,000, and 100,000 iterations.

For the samplers which used the true posterior variance, no attempt was made to exclude burn-in iterations since the samplers were started at posterior modes. For the multi-stage tuned samplers, the iterations during the tuning phase (the first 1/3 of the iterations) were discarded when computing later summary measures. Overall mean squared errors (MSEs) were computed from the univariate summary measures by collapsing across dimensions and simulation runs. Relative efficiency was then calculated as the ratio of the MSE of the sampler of interest over the MSE of the independence proposal.

4.4 Results

Summaries of the four dimensional results are given in tables 3, 4 and 5. In addition, Figure 5 gives representative plots of the MSE for the estimated mean of each sampler as a function of the number of iterations. Summaries of the twenty dimensional results are given in tables 6, 7 and 8. In the tables and figures, we use the following acronyms: "Indep" for the independence sampler, "CNM" for the componentwise Metropolis sampler, "NM" for the random-walk Metropolis sampler, and "NKC" for the Normal Kernel Coupler.

The simulation results show that, as expected, the custom independence proposal gives the lowest MSE when the true covariance of the modes is known. When the true covariance is unknown and must be estimated, the NKC has the best overall performance. The performance benefit of the NKC is more apparent in twenty dimensions than in four and is especially clear for the twenty dimensional multi-modal distributions. Notably, the "tuned" NKC, is more efficient overall than either of the Metropolis samplers constructed using the true covariance of the target distribution. This result holds for both unimodal and bimodal distributions.

When the variance is easy to estimate, as is the case of targets with spherical modes in four dimensions, the independence proposal with the multi-stage tuning method performs at least as well as the NKC. Even in the cases where the tuned independence proposal performed better than the NKC, the performance loss by using the NKC was at most 50%.

The NKC seems to have the greatest performance benefit when the variance of the modes is difficult to estimate. The performance benefit of the NKC over the other tuned methods for more difficult problems was often 500% and sometimes several orders of magnitude.

Taken together, the simulation results show that the NKC outperforms the commonly applied random-walk samplers, across a range of problem types and dimensions even when these use the true posterior variance and the "optimal" scaling. The NKC also outperforms the custom independence sampler when the true mode variance is unknown.

Table 3: All four dimensional distributions: MSE at 10,000 iterations. MSE values have been multiplied by 1×10^5 and are accurate to within $\pm 30\%$. Note that the "Indep" sampler with the "True" variance samples directly from the posterior distribution except in one case (HeavyAndLight).

Sampler	Variance Est.	Mean MSE	2.5% Quantile MSE	97.5% Quantile MSE	Accept Rate
Indep	True	11	121	423	0.95
CNM	True	125,000	314,000	157,900	0.17
NM	True	107,00	33,800	615	0.11
NKC	True	421	155	407	0.44
Indep	3-Stage	1,940	6,710	619	0.74
CNM	3-Stage	125,000	300,000	151,000	0.36
NM	3-Stage	122,000	354,000	181,000	0.54
NKC	3-Stage	176	245	390	0.47

Table 4: Unimodal four dimensional distributions: MSE at 10,000 iterations. MSE values have been multiplied by 1×10^5 and are accurate to within $\pm 30\%$. Note that the "Indep" sampler with the "True" variance samples directly from the posterior distribution.

Sampler	Variance Est.	Mean MSE	2.5% Quantile MSE	97.5% Quantile MSE	Accept Rate
Indep	True	12	2,650	2,450	1.00
CNM	True	14,700	25,700	18,800	0.31
NM	True	840	5,800	5,180	0.21
NKC	True	46	1,320	1,550	0.55
Indep	3-Stage	497	3,320	5,590	0.71
CNM	3-Stage	19,600	29,300	32,300	0.31
NM	3-Stage	4,940	7,430	9,170	0.24
NKC	3-Stage	76	1,620	1,290	0.57

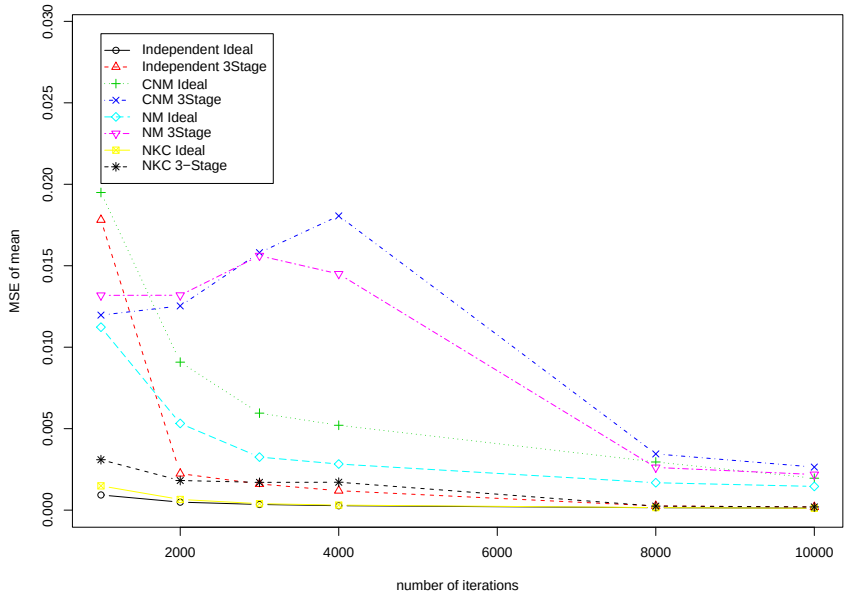
Table 5: Bimodal four dimensional distributions: MSE at 10,000 iterations. MSE values have been multiplied by 1×10^5 and are accurate to within $\pm 30\%$. Note that the "Indep" sampler with the "True" variance samples directly from the posterior distribution except in one case (HeavyAndLight).

Sampler	Variance Est.	Mean MSE	2.5% Quantile MSE	97.5% Quantile MSE	Accept Rate
Indep	True	187	138	5,570	0.91
CNM	True	218,000	5,470,000	2,730,000	0.07
NM	True	18,700	587,000	6,880	0.04
NKC	True	733	1,720	5,960	0.36
Indep	3-Stage	3,370	115,000	6,650	0.76
CNM	3-Stage	217,000	5,220,000	2,610,000	0.40
NM	3-Stage	213,000	6,190,000	3,160,000	0.77
NKC	3-Stage	303	3,080	5,860	0.39

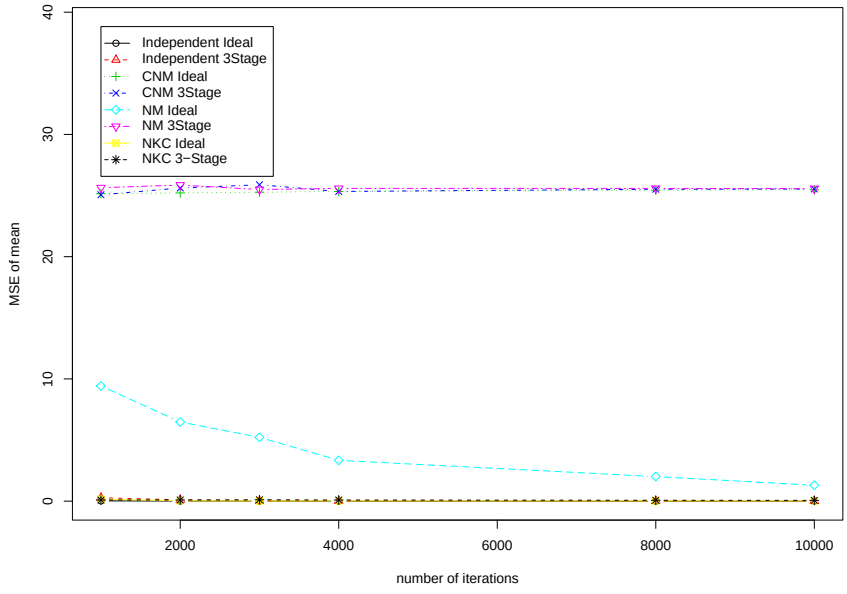
Table 6: All twenty dimensional distributions: MSE at 100,000 iterations. MSE values have been multiplied by 1×10^5 and are accurate to within $\pm 30\%$. Note that the "Indep" sampler with the "True" variance samples directly from the posterior distribution except in one case (HeavyAndLight).

Sampler	Variance Est.	Mean MSE	2.5% Quantile MSE	97.5% Quantile MSE	Accept Rate
Indep	True	10	1,370	4,520	0.95
CNM	True	1,240,000	3,120,000	1,580,000	0.17
NM	True	996,000	2,400,000	1,090,000	0.09
NKC	True	138,000	25,200	18,700	0.04
Indep	3-Stage	1,360,000	1,180,000	3,610,000	0.24
CNM	3-Stage	1,230,000	3,000,000	1,520,000	0.43
NM	3-Stage	1,240,000	2,960,000	1,600,000	0.31
NKC	3-Stage	190,000	281,000	20,000	0.03

Figure 5: MSE for means from the four dimensional simulation. Two representative plots.



(a) "OneMode" Spherical Normal



(b) "HeavyAndLight" One large normal with low probability, and one small normal with high probability

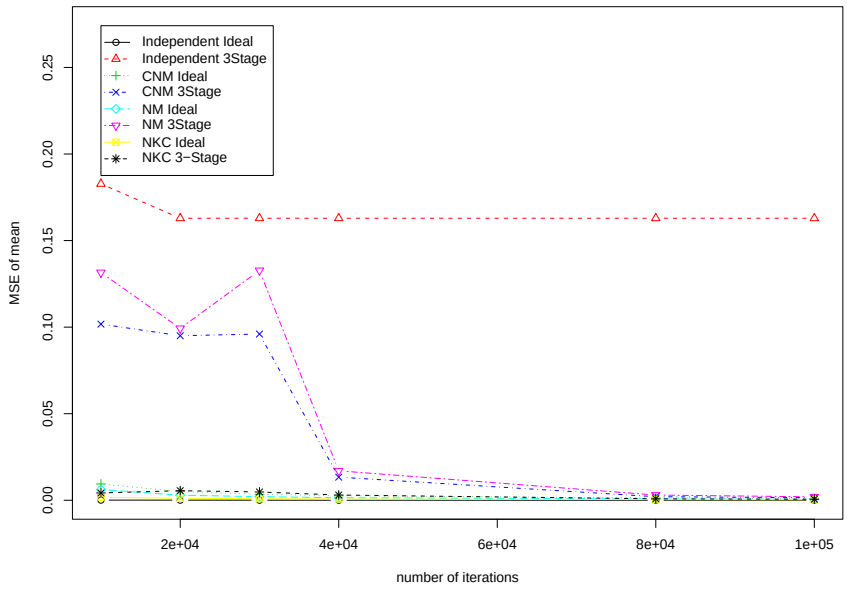
Table 7: Unimodal twenty dimensional distributions: MSE at 100,000 iterations. MSE values have been multiplied by 1×10^5 and are accurate to within $\pm 30\%$. Note that the "Indep" sampler with the "True" variance samples directly from the posterior distribution.

Sampler	Variance Est.	Mean MSE	2.5% Quantile MSE	97.5% Quantile MSE	Accept Rate
Indep	True	1	3,180	3,150	1.00
CNM	True	18,300	44,200	28,600	0.30
NM	True	3,650	10,200	11,800	0.17
NKC	True	58	11,700	11,600	0.06
Indep	3-Stage	8,480	316,000	309,000	0.37
CNM	3-Stage	16,900	40,100	30,000	0.41
NM	3-Stage	5,710	47,300	39,000	0.30
NKC	3-Stage	1,880	17,700	14,300	0.05

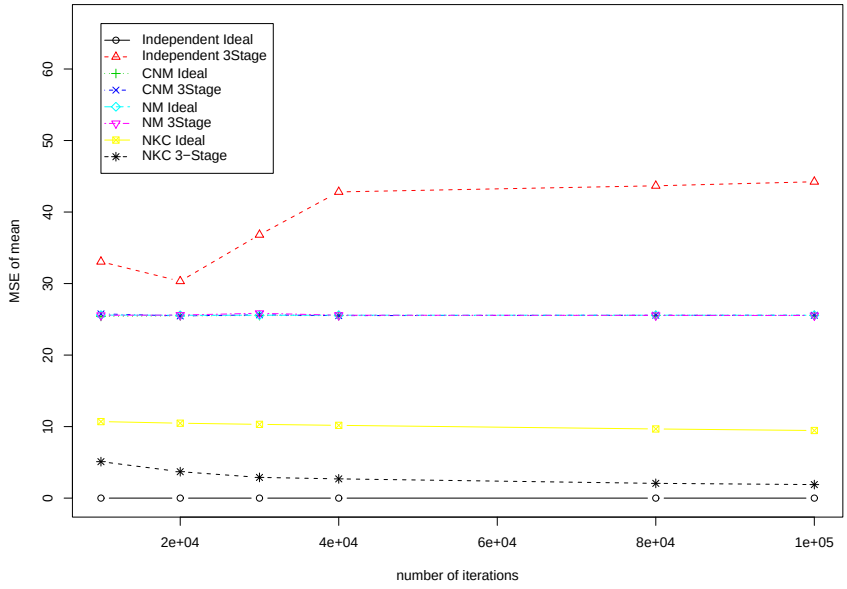
Table 8: Bimodal twenty dimensional distributions: MSE at 100,000 iterations. MSE values have been multiplied by 1×10^5 and are accurate to within $\pm 30\%$. Note that the "Indep" sampler with the "True" variance samples directly from the posterior distribution except in one case (HeavyAndLight).

Sampler	Variance Est.	Mean MSE	2.5% Quantile MSE	97.5% Quantile MSE	Accept Rate
Indep	True	16	13	5,550	0.91
CNM	True	2,160,000	5,430,000	2,740,000	0.07
NM	True	1,740,000	4,190,000	1,900,000	0.03
NKC	True	241,000	35,300	24,100	0.03
Indep	3-Stage	2,380,000	1,820,000	6,090,000	0.13
CNM	3-Stage	2,140,000	5,210,000	2,630,000	0.45
NM	3-Stage	2,160,000	5,150,000	2,770,000	0.31
NKC	3-Stage	331,000	478,000	24,200	0.02

Figure 6: MSE for means from twenty dimensional simulation. Two representative plots.



(a) "OneMode" Spherical Normal



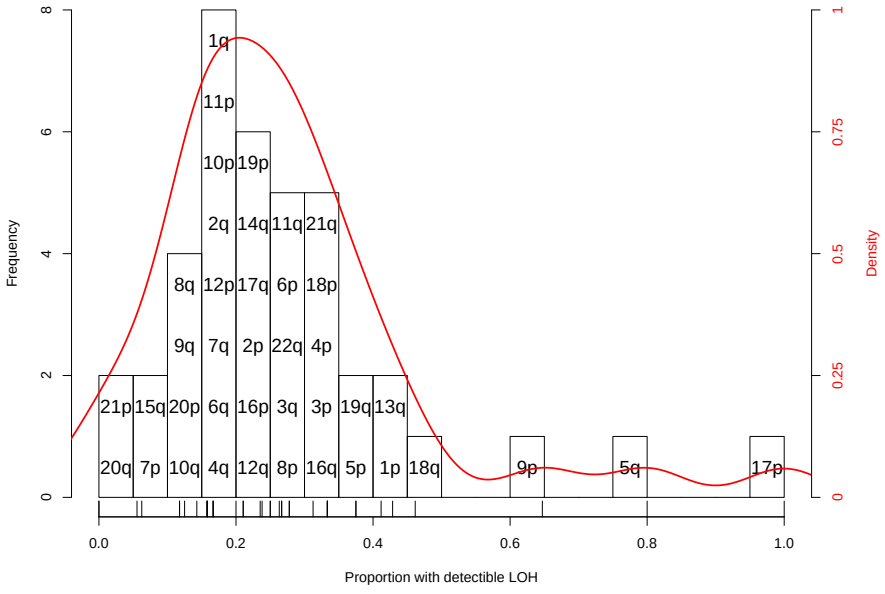
(b) "HeavyAndLight" One large normal with low probability, and one small normal with high probability

5 Application: Genetic Instability of Esophageal Cancers

Cancer cells undergo a number of genetic changes during neoplastic progression, including loss of entire chromosome sections. When an individual patient has two different alleles for a particular gene, the loss of a chromosome section containing one allele by abnormal cells, termed "Loss of Heterozygosity" (LOH), can be detected using laboratory assays. Chromosome regions with high rates of LOH are hypothesized to contain genes which regulate cell behavior so that loss of these regions disables important cellular controls.

The Seattle Barrett's Esophagus research project (Barrett et al., 1996) has collected LOH rates from esophageal cancers for 40 regions, each on a distinct chromosome arm. The intent is to locate "Tumor Suppressor Genes" (TSGs), whose deactivation contributes to the development of esophageal cancer (Fearon, 1998; Klein, 1987). Chromosome regions with high rates of LOH ("systematic LOH") are hypothesized to contain TSGs, (Newton et al., 1998; Marshall, 1991). In addition to LOH of regions containing TSGs, there is also a high

Figure 7: Histogram (bars and left axis) and kernel density estimate (curve and right axis) for the Barrett's LOH data. Text labels give the location of each chromosome arm.



level of "background" LOH which is thought to be a consequence, rather than a cause, of neoplastic progression. A histogram of the relative frequency of LOH for the Barrett's data is shown in Figure 7.

The immediate goal of this analysis is to determine the probability of LOH for both the "background" and TSG groups. This will enable the development of a simple discrimination method. Since the labeling of the two groups is unknown, we model the LOH frequency using mixture models, as described by Desai (2000). While several models have been considered, we will focus on a hierarchical Binomial-BetaBinomial mixture model:

$$\begin{aligned} X_i &\sim \eta \text{ Binomial}(N_i, \pi_1) \\ &\quad + (1-\eta) \text{ Beta-Binomial}(N_i, \pi_2, \gamma) \\ \eta &\sim \text{Unif}[0,1] \\ \pi_1 &\sim \text{Unif}[0,1] \\ \pi_2 &\sim \text{Unif}[0,1] \\ \gamma &\sim \text{Unif}[-30,30] \end{aligned}$$

where η is the probability of a location being a member of the binomial group, π_1 is the probability of LOH in the binomial group, π_2 is the probability of LOH in the beta-binomial group, and γ controls the variability of the beta-binomial group (on the logit scale)

We have parameterized the Beta-Binomial so that γ_2 is a variance parameter defined on the range $-\infty \leq \gamma_2 \leq \infty$. As $\gamma_2 \to -\infty$ the beta-binomial becomes a binomial and as $\gamma_2 \to +\infty$ the beta-binomial becomes a uniform distribution on $[0,1]$. This results in the unnormalized posterior density

$$p(\eta, \pi_1, \pi_2, \gamma) = \prod_{i=1}^{N} f(x_i, n_i | \eta, \pi_1, \pi_2, \omega_2) \tag{2}$$

on the prior range, where

$$\begin{aligned} f(x, n | \eta, \pi_1, \pi_2, \omega_2) &= \eta \binom{n}{x} \pi_1^x (1-\pi_1)^{n-x} \\ &\quad + (1-\eta) \binom{n}{x} \frac{\Gamma(\frac{1}{\omega_2})}{\Gamma(\frac{\pi_2}{\omega_2})\Gamma(\frac{1-\pi_2}{\omega_2})} \frac{\Gamma(x+\frac{\pi_2}{\omega_2})}{\Gamma(n-x+\frac{1-\pi_2}{\omega_2})\Gamma(n+\frac{1}{\omega_2})} \end{aligned} \tag{3}$$

and $\omega_2 = \frac{\exp(\gamma)}{2(1+\exp(\gamma))}$.

Unlike most mixture models, where all of the components come from the same parametric family, the proposed model mixes two different distributions; a binomial (with one parameter) and a beta-binomial (with two parameters). We have intentionally omitted fixing which mixture component corresponds to the background group and which corresponds to the TSG group so that we can discover which of the two possible arrangements is better supported by the data. As a consequence, the model has two well separated *non-symmetric* modes, both of which may contribute considerable probability mass. Thus, accurate estimation of this posterior density requires effective sampling from both modes.

5.1 Fitting

To locate posterior modes, we employed the Nelder-Mead function maximizer provided with the software package `R` (Ihaka & Gentleman, 1996). We started the maximizer from a large number of initial states sampled from the prior. This yielded two well separated, peaks one at (0.903, 0.228, 0.708, 3.54) and another at (0.078, 0.832, 0.230, -18.51). The first mode has log-likelihood of -88.09, while the second is somewhat lower at -90.01. In the absence of other information, we would expect the modes to have posterior probability of approximately $0.87 = 1/(1 + \exp(-90.01 - (-88.09)))$ and $0.13 = 1 - 0.87$, respectively.

We constructed the NKC using 120 component states and initialized half of the states to each of the two local maxima. Starting with the prior variance, we used two preliminary runs

Table 9: Parameters and likelihood maxima for the Binomial-BetaBinomial model.

		Likelihood Maxima	
Parameter	Description	Mode 1	Mode 2
η	Proportion in Group 1	0.903	0.078
π_1	Group 1 probability of LOH	0.228	0.832
π_2	Group 2 probability of LOH	0.708	0.230
γ_2	Variability of LOH in Group 2	3.54	-18.51
	Log-Likelihood	-88.09	-90.01

of 6,480 iterations (54 complete scans) to estimate the variance of the two modes. We then ran the NKC for 23,640 iterations (197 complete scans) to generate samples for estimation. (See the Warnes (2000b) for a complete description of the fitting process.) Using the output from the NKC, we estimated posterior means and credible regions for the entire posterior distribution as well as for the individual modes.

5.2 Results

Using the output from the NKC, we estimated posterior means and credible regions for the entire posterior distribution as well as for the individual modes. Table 10 gives these estimates, as well estimates obtained by direct numerical integration using adaptive quadrature (Berntsen et al., 1991).

Figure 8 gives a 2 dimensional histogram constructed from the output of simulation 3

Figure 8: 2-d histogram of the joint marginal density of π_1 and π_2 generated from the MCMC simulation. Bins are boxes of side length 0.01 and intensity is proportional to log-frequency.

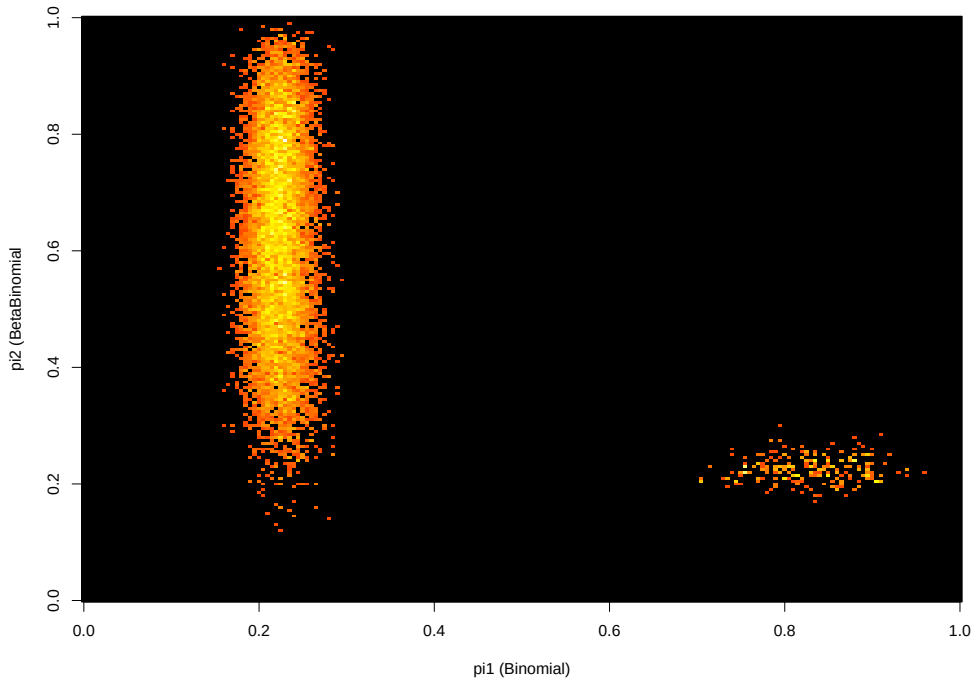


Table 10: Means and 95% credible intervals for the Binomial-BetaBinomial model

Overall Estimates

	Adaptive Quadrature	Estimates		
		Mean	2.5% Quantile	97.5% Quantile
η	0.832	0.82	0.0748	0.965
π_1	0.246	0.257	0.193	0.829
π_2	0.617	0.612	0.23	0.912
γ_2	12.82	12.3	-21.2	29.3
Prob(Mode 1)	0.970	0.954		
Prob(Mode 2)	0.030	0.047		

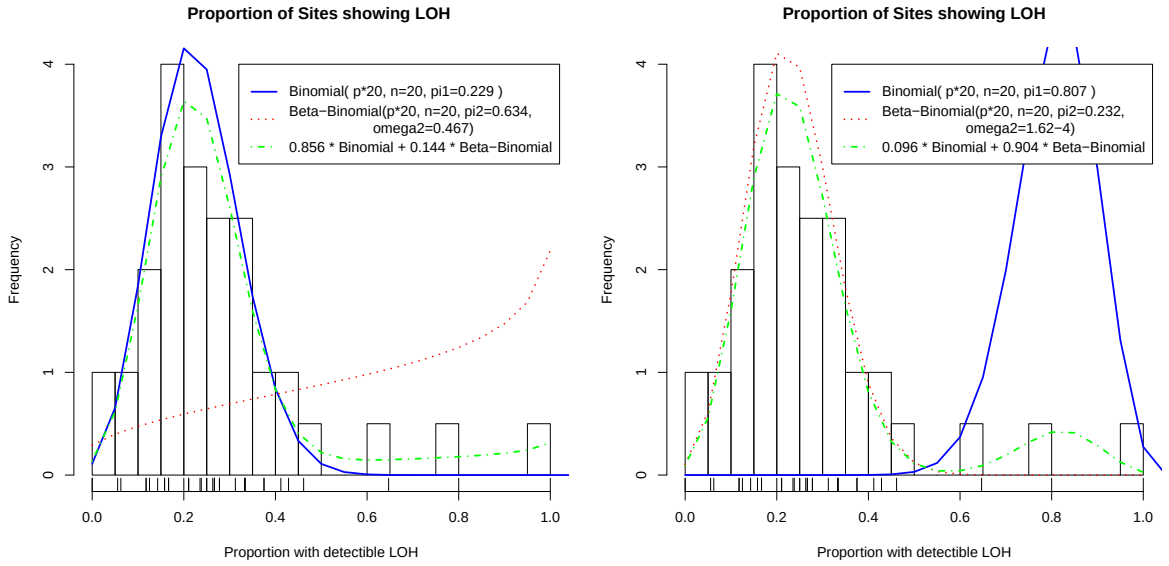
Mode 1 Estimates

	Adaptive Quadrature	Estimates		
		Mean	2.5% Quantile	97.5% Quantile
η	0.854	0.856	0.656	0.966
π_1	0.229	0.229	0.192	0.266
π_2	0.629	0.631	0.318	0.913
γ_2	13.73	13.7	-4.97	-29.3

Mode 2 Estimates

	Adaptive Quadrature	Estimates		
		Mean	2.5% Quantile	97.5% Quantile
η	0.091	0.0839	0.0174	0.219
π_1	0.825	0.832	0.741	0.914
π_2	0.232	0.23	0.199	0.261
γ_2	-16.28	-17.5	-29.5	-4.11

Figure 9: Histogram and fitted distributions (curves) for the Barrett's LOH data



(a) Binomial-BetaBinomial Mode

(b) Binomial-Binomial Mode

clearly showing the asymmetry of the two modes. We estimate that smaller mode contains only 4.7% of the posterior mass. For this mode, the variance parameter γ is very small (-17.5), forcing the BetaBinomial component to act like a Binomial. The fact that the preferred mode uses the Binomial mixture to explain the background loss rate and that the secondary mode forces the BetaBinomial to act like a Binomial with the same mean suggests that a binomial model is sufficient for the "background" probability of LOH.

While both modes estimate the LOH probability in the "background" group to be approximately 0.23, the two modes give quite different distributions for the TSG group. Figures 10(a) and 10(b) plot the fitted distributions for each mode against a histogram of the original data. Figure 10(b) shows that the "Binomial-Binomial" mode assigns most of the mass for the TSG group to a binomial which has its density concentrated near 0.83. In contrast, Figure 10(a) shows that the larger "Binomial-BetaBinomial" mode, which contains roughly 96% of the posterior probability, spreads the TSG group much wider, with considerable probability density over the range of the "background" group.

Figure 10: Posterior probability (curve, right axis) of membership in the TSG group. The histogram (left axis) shows the observed LOH rate with text labels giving the location of each chromosome arm. Chromosome arms with 50% or higher posterior probability of membership in the TSG group are labeled in red.

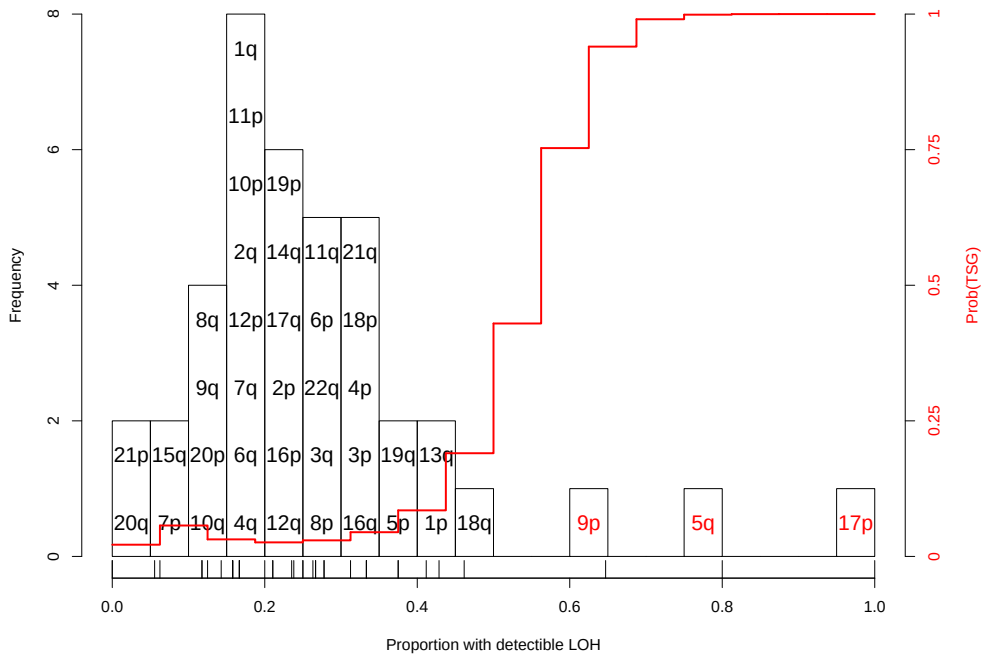


Figure 10 plots the estimated posterior probability of belonging to the TSG group as a function of the LOH rate. Regions with LOH frequencies above 50% almost certainly belong to the TSG group, while regions with loss rates below 50% are more likely to belong to the background group, although there is still a possibility that they belong instead to TSG group.

Only five chromosome arms have a TSG group membership probability above 10%: 1p, 13q, 9p, 5q and 17p. Of these, only 9p, 5q, and 17p, with memberships probabilities 0.975, >0.999, and >0.999, are more likely to belong to the TSG group than the background group. This aligns well with research on the biology of TSGs and the role of LOH in the inactivation of these genes.

In particular, 17p is the location of the p53 (TP53) gene and 9p is the location of the p16

Table 11: Chromosome Arms with more than 10% posterior probability of belonging to the TSG group.

Chromosome Arm	Observed Prob(LOH)	Binomial Quantile	Posterior Prob(TSG)
1p	0.41	0.98	0.14
13q	0.43	0.98	0.15
18q	0.46	0.99	0.21
9p	0.65	> 0.99	0.98
5q	0.80	> 0.99	> 0.99
17p	1.00	> 0.99	> 0.99

(CDKN2A) gene. Barrett et al. (1999) showed that LOH of 17p and 9p, along with mutation or hyper-methylation of the remaining p53 or p16 allele is necessary for the development of esophageal cancer from Barrett's epithelium. Barrett et al. (1999) also evaluated the role of LOH at 5q, 13q, and 18q in the development of cancer. For these sites they found no evidence that LOH was required for the development of cancer.

6 Discussion

In this text, we have introduced the Normal Kernel Coupler, a conceptually simple method for sampling from posterior distributions that can be applied whether the target distribution has one or several modes. We have proven that the NKC is a ergodic MCMC sampler for any continuous distribution on d-dimensional Euclidean space. We have also shown that the NKC outperforms standard random-walk Metropolis samplers and a custom independence sampler when the true variance is unknown. In fact, the NKC outperforms the random-walk Metropolis samplers constructed using the true posterior variance. We have demonstrated these methods on a real example using a model with two distinct and dissimilar modes. The results from fitting this model using the NKC compare favorably with with those obtained by adaptive quadrature at much lower computational cost.

The current implementation of the NKC is inefficient in high dimensions. In part, this

is a consequence of updating all of the parameters of a given component state as a block. While it would be possible to update only a small subset of the parameters, this is not a reasonable approach in the context of multiple modes because it prevents moves between unconnected modes that do not lie along the coordinate axes. Instead, we are exploring an approach similar to the Adaptive Direction Sampler (Gilks et al., 1994), where updates are generated on a subspace selected adaptively using the set of current states.

Another method of improving the performance of the NKC in high dimensions is to "retry" failed proposals. The idea, introduced by Tierney & Mira (1999), is to generate a second candidate point from a different proposal distribution if the initial proposal is rejected. This second candidate point is then accepted or rejected using an adjusted acceptance function. By this means, when a NKC proposal is rejected, a standard random-walk Metropolis step can be performed instead. This would increase the overall acceptance rate, and would allow for more local moves than otherwise possible.

The flexibility and performance of the NKC on a variety of unimodal and multi-modal distributions makes it a promising tool for sampling from multi-modal distributions in low and moderate dimensions. Software implementing the NKC is available as part of the `Hydra` MCMC library (Warnes, 2000b; Warnes, 2001) developed by the author and available free of charge from the author's web site (Warnes, 2000a).

References

Barrett, M. T., Galipeau, P. C., Sanchez, C. A., Emond, M. J., & Reid, B. J. (1996). Determination of the frequency of loss of heterozygosity in esophageal adenocarcinoma by cell sorting, whole genome amplification and microsatellite polymorphisms. *Oncogene* **12**, 1873–1878.

Barrett, M. T., Sanchez, C. A., Prevo, L. J., Wong, D. J., Paulson, T. G., Rabinovich, P. S., & Reid, B. J. (1999). Evolution of neoplastic cell lineages in Barrett oesophagus. *Nature Genetics* **22**, 106–109.

Berntsen, J., Espelid, T. O., & Genz, A. (1991). An adaptive multidimensional integration routine for a vector of integrals. *Transactions on Mathematical Software* **17**, 452–456.

Chan, K. S. & Geyer, C. J. (1994). Comment on "Markov chains for exploring posterior distributions". *The Annals of Statistics* **22**, 1747–1758.

Desai, M. (2000). *Mixture Models for Genetic Changes in Cancer Cells.* PhD thesis, University of Washington.

Fearon, E. R. (1998). Tumor suppressor genes. *The Genetic Basis of Human Cancer* **7**, 145.

Gelman, A., Roberts, G., & Gilks, W. (1995). Efficient Metropolis jumping rules. In Berger, J. O., Bernardo, J. M., Dawid, A. P., & Smith, A. F. M., editors, *Bayesian Statistics V*. Oxford University Press.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 473–483.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In Keramidas, E. M., editor, *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, pages 156–163, Seattle, WA. Interface Foundation of North America.

Geyer, C. J. & Thompson, E. A. (1994). Annealing Markov chain Monte Carlo with applications to ancestral inference. Technical Report 589, School of Statistics, University of Minnesota.

Gilks, W. R., Roberts, G. O., & George, E. I. (1994). Adaptive direction sampling. *The Statistician* **43**, 179–189.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314.

Klein, G. (1987). The approaching era of the tumor suppressor genes. *Science* **238**, 1539–1544.

Mariani, E. & Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters* **19**, 451–458.

Marshall, C. J. (1991). Tumor suppressor genes. *Cell* **64**, 313–326.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., & Teller, A. H. (1953). Equations of state calculations by fast computing machine. *Journal of Chemical Physics* **21**, 1087–1091.

Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing* **6**, 353–366.

Newton, M. A., Gould, M. N., Reznikoff, C. A., & Haag, J. D. (1998). On the statistical analysis of allelic–loss data. *Statistics in Medicine* **17**, 1425–1445.

Raftery, A. E. & Lewis, S. M. (1996). Implementing MCMC. In *Markov Chain Monte Carlo in Practice*, pages 115–130. Chapman & Hall.

Tierney, L. (1996). Introduction to general state–space markov chain theory. In *Markov Chain Monte Carlo in Practice*, pages 59–74. Chapman & Hall.

Tierney, L. & Mira, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine* **18**, 2507–2515.

Tjielmeland, H. & Hegstad, B. K. (2000). Mode jumping proposals in MCMC. *Scandinavian Journal of Statistics* (to appear).

WARNES, G. R. (2000a). Hydra mcmc web site. Web Site. `http://www.warnes.net/MCMC`.

WARNES, G. R. (2000b). *The Normal Kernel Coupler: An adaptive Markov Chain Monte Carlo method for efficiently sampling from multi-modal distributions.* PhD thesis, University of Washington.

WARNES, G. R. (2001). Hydra: A java library for markov chain monte carlo. Technical Report 394, Deprtment of Statistics, University of Washington.

A Proof of Theorem 1

Proof 1

1. $\pi_*^{(\cdot)}$-irreducible

   The transition kernel P for the NKC from a point x to a set A is

   $$P(x, A) = \int_A q_*(x, y)\alpha(x, y)\mu(dy),$$

   and the n-step transition kernel is defined recursively by

   $$P^n(x, A) = \int P(x, dy)P^{(n-1)}(y, A)$$

   for $n \geq 2$ where $P(x, dy)$ is the probability of moving to a small measurable subset $dy \subset S$ given that the move starts at x.

   We need to show that for all $x \in \Re^{d\times C}$ and $A \subset \Re^{d\times C}$ there is a value of n for which

   $$P^n(x, A) > 0 \text{ for whenever } \pi_*(A) > 0. \tag{4}$$

   First, note that the conditions on $h^2\mathbf{V}$ guarantee that

   $$q_k(Y^{(i)}|X^{(\cdot)}) = \sum_{i=1}^{C} \mathrm{N}_d\big(Y^{(i)}|X^{(j)},\ h^2\mathbf{V}\big) > 0 \qquad \text{for all } Y^{(i)} \in \Re^d \text{ and } X^{(\cdot)} \in \Re^{d\times C}\ . \tag{5}$$

Consequently,

$$\alpha(X^{(i)}, Y^{(i)}|X^{(-i)}) = \min\left\{1, \frac{p(Y^{(i)})\ q_k(Y^{(i)} \to X_t^{(i)}|X_t^{(-i)})}{p(X^{(i)})\ q_k(X_t^{(i)} \to Y^{(i)}|X_t^{(-i)})}\right\} > 0 \tag{6}$$

whenever $p(Y^{(i)}) \propto \pi(Y^{(i)}) > 0$.

Let $P(X^{(i)}, Y^{(i)}|X^{(\cdot)})$ be the transition kernel for a single step of the NKC which updates component i conditional the set of current states. Without loss of generality, assume that the permutation defining the order of component updates is $1, 2, .., C$. Now, the transition probability for one complete scan of the C component states is

$$\begin{aligned}
P^C(X^{(\cdot)}, Y^{(\cdot)}) &= \prod_{i=1}^{C} P(X^{(i)}, Y^{(i)}|(Y^{(1,..i-1)}, X^{(i,..,C)})) && (7)\\
&= \prod_{i=1}^{C} \{\, q_k(Y^{(i)}|(Y^{(1,..i-1)}, X^{(i,..,C)})) \times && (8)\\
&\qquad \alpha(X^{(i)}, Y^{(i)}|(Y^{(1,..i-1)}, X^{(i,..,C)}))\,\} && (9)\\
&> 0 \quad \text{for all } X^{(\cdot)}, Y^{(\cdot)} \in \Re^{d\times C} \text{ whenever } \pi_*(X^{(\cdot)}) > 0 && (10)
\end{aligned}$$

by (5) and (6). Since this holds for any $Y^{(\cdot)} \in A \subset \Re^{d\times C}$ the NKC is irreducible for π_*.

2. Harris recurrent

Chan & Geyer (1994) give sufficient conditions for an irreducible variable-at-a-time Metropolis-Hastings sampler to be Harris recurrent:

Theorem 3 *A variable-at-a-time Metropolis-Hastings algorithm on R^d with proposal distributions that are absolutely continuous with respect to Lebesgue measure is Harris recurrent if all of the conditional samplers (including the unconditional sampler which conditions on the empty set of variables $I = \emptyset$) are irreducible for any values of the fixed variables.*

The essence of the Chan-Geyer condition is that the sampler is Harris recurrent if updates on any subset of the components $I = \{i_1, .., i_k\}$, conditional on the remaining components $I^- = \{i_{k+1}, .., i_C\}$, are irreducible for any fixed values of the components I^-. See Chan & Geyer (1994) for the proof.

For the NKC, the Chain-Geyer condition is easily verified. Consider the k-step transition kernel for the conditional sampler with $C - k$ components held fixed:

$$\begin{aligned} P^k(X^{(I)}, Y^{(I)} | X^{(I^-)}) &= \prod_{i=1}^{k} P(X^{(i)}, Y^{(i)} | (Y^{(1,..i-1)}, X^{(i,..,k)}), X^{(-I)}) \\ &= \prod_{i=1}^{k} P(X^{(i)}, Y^{(i)} | (Y^{(1,..i-1)}, X^{(i,..,C)})) \\ &= \prod_{i=1}^{k} \{\, q_k(Y^{(i)} | (Y^{(1,..i-1)}, X^{(i,..,C)})) \times \\ &\qquad \alpha(X^{(i)}, Y^{(i)} | (Y^{(1,..i-1)}, X^{(i,..,C)})) \,\} \\ &> 0 \quad \text{for all } X^{(I)}, Y^{(I)} \in \Re^{d \times k} \text{ and } X^{(-I)} \in \Re^{d \times C - k} \\ &\qquad \text{whenever } \pi_*(X^{(\cdot)}) > 0. \end{aligned}$$

Consequently, each of the conditional samplers is irreducible, the requirements of theorem 3 are met, and the NKC is Harris recurrent. Since the chain has invariant distribution π_* by construction, the conditions to Tierney (1996) Theorem 4.2 and 4.3 are met. Consequently, the NKC has π_* as its unique invariant distribution, it converges to this distribution, and sample path averages computed using NKC converge to the true expectation under π_*.

3. Aperiodic

Recall that an m-cycle for an irreducible chain with transition kernel P is a collection $\{E_0, E_1, \ldots, E_{m-1}\}$ of disjoint sets such that

$$P(x, E_j) = 1 \text{ for } j = i + 1 \mod m \text{ and for all } x \in E_i.$$

Define the support S of π_* as $S = \{x \in \Re^{d \times C} : \pi_*(x) > 0\}$. Without loss of generality, we can assume that $E_i \subset S$ for $i = 0, \ldots, m-1$. Now, if an m-cycle is present, then

$$P^t(x, E_k) = 1 \text{ for } k = i + t \mod m \text{ and for all } x \in E_i.$$

Consider $t = C$. We have previously shown that $P^C(x, A) > 0$ for all $x \in \Re^{d \times C}$ and $A \in S$. Consequently $P^C(x, E_k) = 1$ for all $X \in E_i$ can only hold if $m = 1$ so that $E_0 = E_k = S$. Since the largest cycle length is 1, the NKC is aperiodic.